Multilingual Ontology of Proper Names

Cvetana Krstev, Duško Vitas, Denis Maurel, Mickaël Tran

Multilingual Ontology of Proper Names

Cvetana KRSTEV (1), Duško VITAS (2), Denis MAUREL (3), Mickaël TRAN (3)

(1) Faculty of Philology, University of Belgrade
Studentski trg 3, CS – Belgrade, cvetana@matf.bg.ac.yu
(2) Faculty of Mathematics, University of Belgrade
Studentski trg 16, CS – Belgrade, vitas@matf.bg.ac.yu
(3) Université François-Rabelais de Tours, LI
EPU-DI, 64 avenue Jean-Portalis, 37200 Tours, France
denis.maurel@univ-tours.fr, mickael.tran@etu.univ-tours.fr

Abstract

This paper deals with a multilingual four-layered ontology of proper names. This ontology is organized around a conceptual proper name that represents the same concepts in different languages.

1. Introduction

Natural language processing depends on the available language resources, adapted to envisaged methods and applications. A considerable number of applications use a resource that consists of a word list furnished with morphological, syntactic, and semantic information, which is known as an *electronic dictionary* (abbr. *e-dictionary*). The scope and use of an e-dictionary depend on the model of language that is used: one that combines the stochastic and linguistic information, or one that relies on linguistic information only. Among the e-dictionaries of the second type, dictionaries of common names, such as dictionary system DELA (Courtois and Silberztein 1990), or special terminological resources (Sager 1990), are dominant.

One specific category of words, *proper names*, is often neglected when constructing an e-dictionary. Proper names, however, constitute a significant part of many texts, and therefore, the text analysis based on lexical recognition can encounter serious problems. For instance, proper names represent more then 10% of a newspaper text (Coates-Stephans 1993), and also a considerable part of "unknown words" in a corpus. Furthermore, in a novel the names of main characters only can represent more than 1% of all running words.

2. The Problem

The lack of an exhaustive description of proper names that would include the precise description of their linguistic properties is prominent. The reasons for this are twofold. First, different heuristic procedures reduce the extraction of proper names to more or less successful approximations, such as:

a) An approximation using the rule with the minimal list of proper names since "through a judicious use of internal and external evidence relatively small gazetteers are sufficient to give good Precision and Recall" (Mikheev et al. 1999);
b) An approximation that relies on the orthographic practice that proper names are written using the initial capital letters. From this a very simple criteria can be deduced that "a proper name is any unknown name that starts with a capital letter". However, this criteria, according to (Maurel 2004), due to the homography and compounds is correct in no more than 50% of cases.

On the other hand, proper names have a special linguistic status since they cannot be subjected to the usual defining methods (such as, answering questions of the type "What does it mean?"). However, minute analysis shows that the proper names are distinguished in one language system by the richness of specific semantic information, so that their systematic, exhaustive and explicit description represents a hard task.

The exhaustive description of proper names becomes even more complex if this description has to be done in the environment of multilingual applications. Although proper names represent cognates par excellence, in many cases their graphic variations can disable their recognition by the approximative string pattern matching. For instance, the pope's name is *Giovanni Paolo II* in Italian, *Jean Paul II* in French, *Juan Pablo II* in Spanish, *John Paul II* in English, *Jovan Pavle II* and *Јован Павле II* in Serbian (using Latin and Cyrillic alphabet), *Ivan Pavao II* in Croatian, etc.

Even in cases when it is possible to spot an occurence of a proper name in a monolingual text with the mentioned simple methods, the identification of a concept that is represented by that name in a multilingual text is possible only exceptionally. It has to be added that proper names share the morphological (derivational and inflectional) properties of languages in which they are realized. One illustrative example is given by possessive and relational adjectives: a phrase *the Chopin tradition* can vary in English as *Chopinian tradition* or *Chopin's tradition*, in Polish it is *Chopinowska tradycja*, in French *la tradition chopinienne*, in Serbian *šopenovska tradicija*. This example also shows that the proper names are not always written with the initial capital letter. The identification of the equivalent proper names becomes more complex if different coding schemes are taken into consideration, particularly Unicode. As the overlapping of the coding schemes is not defined, the identification of, for instance, Latin and Cyrillic representations of same proper names for languages that use both alphabets is not straightforward.

3. Multilingual Ontology of Proper Names

In a multilingual application, a description of proper names cannot be reduced to the construction of a multilingual e-dictionary, due to the complexity of semantic relations that connect them. It seems that in multilingual context it is more suitable to represent proper names as ontology in the sense of (Gruber 1995). The analysis of proper name properties shows that such an ontology must have at least four levels: two language independent levels: conceptual and metaconceptual, and two language dependent levels: linguistic level and instance (Grass et al. 2002). The architecture of such an ontology is represented in Figure 1.

The conceptual level is organized around the *pivot* (the *conceptual proper name*), which is represented by the unique identification number (ID). This has the role of an inter-lingual identifier, enableing the connection of proper names that represent the same concepts in different languages. Conceptual proper names do not correspond directly to the language referents; however, they enable a definition of some relations on the conceptual level, such as synonymy, meronymy, predication. An example of synonymy in a diachronic register is *Zaire*, which has been renamed to *Democratic Republic of Congo* after a coup. *France* and *French Republic* are synonymous only in political context. *Paris* $\subset$ *France* $\subset$ *Europe* represents a relation of meronymy. Predication is a relation that can be established between two proper names using the common name predicates, for instance, *Paris is the capital of France*, *Mozart is the composer of the Magic Flute*. The relation of predication, inspired by Mel'čuk's lexical function *Cap* (Mel'čuk 1984, 1988, 1992), enables recognition of a proper name on the basis of one expansion and local grammars (Gross 1997). Predication is common for represented languages, while expansion and local grammars are specific for each particular language. In the above examples, *the capital of France* and *the composer of the Magic Flute* are anaphora of the associated proper names, namely *Paris* and *Mozart* respectively.

On this level the relationship is established with *WordNets* for which the concept of Inter-Lingual Index (ILI) was introduced for the first time in the scope of the EuroWordNet project (Vossen 1998). Due to this relationship the position of a proper name in the lexical hierarchy imposed by English *WordNet*, and propagated to the *WordNets* of a number of other European languages, can be defined. Thus the concept *Paris*, is represented in English *WordNet* by a synset, a set of near synonyms, <Paris, City of Light, French capital, capital of France>, and the value of its ILI is 0558236-n. Its position in the hypernym hierarchy is:

entity
  location
    region
      area, country
        center, middle, heart
          seat
            capital
              national capital
                Paris, City of Light, ...

The metaconceptual level enables a homogenous classification of proper names on the bases of super-type and type that are associated to every proper name, where supertype classifies proper names according to their traditional syntactic and semantic properties, while type gives a more refined classification of a super-type. For instance, for a supertype *toponym* (location), types can be *astronym, geonym*, *hydronym*, etc. We distinguish also between historical, religious and ficticious names (essence).

Linguistic level describes the realizations of a proper name in the observed language. On this level the canonic forms, or prolexemes, are defined and are connected to the ID for the particular language. The prolexeme for the capital of *France* is *Paris* in French and English, *Pariz* and *Париз* in Serbian, *Paryz* in Polish, *and Париж* in Russian. The aliases are connected to prolexemes that describe the variations in orthography, abbreviated forms, acronyms, etc. For instance, in English aliases for *George Walter Bush* are *George W. Bush, George Bush, Bush*, while aliases for *Ivo Andric* are *Ivo Andrich,* and *Ivo Andrics* etc. Also, in a toponym *Ljubljana* the consonant group *lj* can be recorded as a digraph and as single character (The Unicode Standard, Version 4.0, Latin Extended-B). On this level the relations between a prolexeme and its derivational forms, as well as relations between its aliases and their derivational forms, are established. The examples of this type of relation are the names of masculine and feminine inhabitants of toponyms, possessive and relational adjectives derived from toponyms and inhabitants, etc. For instance, in English *Parisian* is an inhabitant of *Paris*, while in Serbian, *Parižanin* is a masculine inhabitant of *Pariz* (with alias *Parizlija*) and *Parižanka* is a feminine inhabitant of *Pariz*.

Beside the relation of expansion that assigns to a proper name a local grammar, viewed as a generalization of a synset, on a linguistic level other relations are defined, such as:

a) A Blark (Basic LAnguage Ressources Kit) relation, which connects a proper name to some time period or region in order to indicate its relevance in that period or region (Cucchiarini et al. 2000).
b) Antonomasia, by which a proper name becomes a common name. For instance, in French the proper name *kleenex* has become a common name for a paper handkerchief, in English *biro* for a ball point pen, in Serbian *žilet* (transcribed from *Gillette*), has become a common name for a razor blade.
c) A sorting relation gives the information on how to classify multiword proper names. Namely, many dictionaries arrange multiword proper names by inverting their constituent parts (Tran et al. 2005).

The **level of instances** contains the inflected forms of proper names that are linguistically described (for instance, their inflectional properties are given). The relation between lemmas on a linguistic level and their forms on the level of instances can be defined by the code of the inflectional class. For many European languages this code corresponds to the code assigned to each lemma in the DELA-type dictionary.

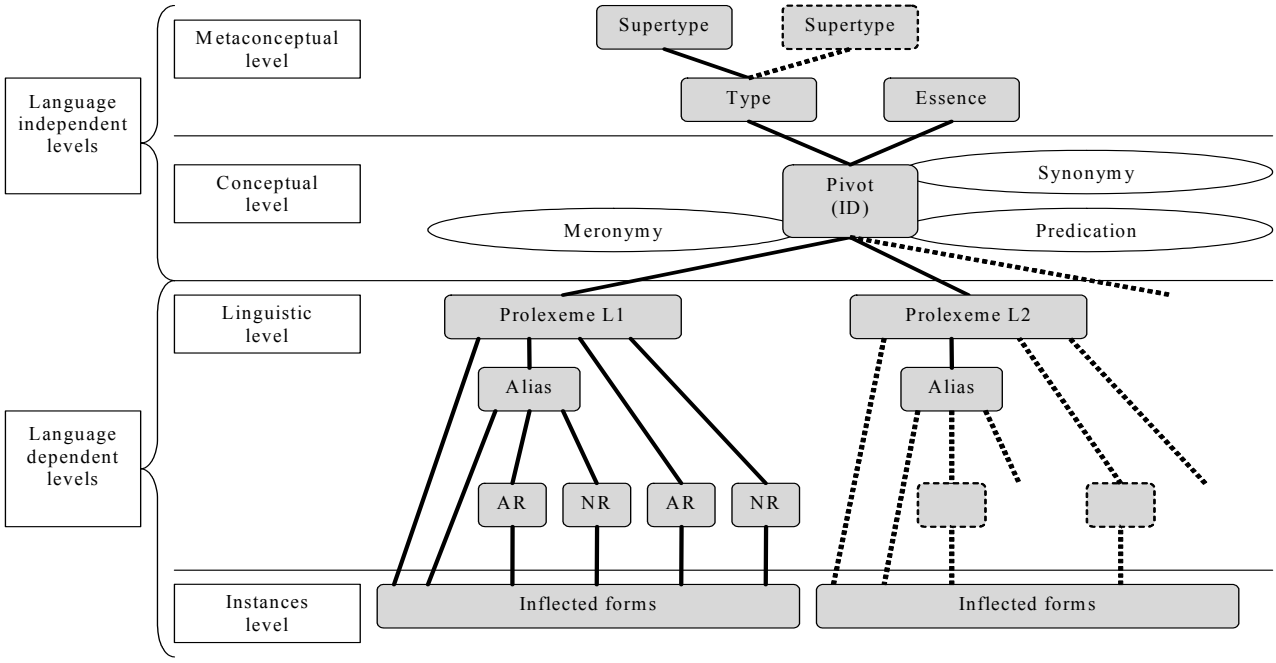

Figure 1. The general structure of the ontology of proper names

For instance, in Serbian the sex of a person, as well as the order of his/her first name and surname, influence the inflection of personal name. Also, derivational properties of multiword proper names, such as derivation of possessive and relational adjectives, still remain to be described.

4. One example

As an example of the multilingual ontology of proper names, the simplified model of the implementation of the proper name *Paris* in French and in Serbian is represented in Figure 2.

On the conceptual level, the unique pivot (conceptual proper name) corresponds to the proper name *Paris*, represented with the value 3200 of the identifier ID. The relation ILI connects it to the *WordNet*, and thus specifies its position in the hypernym hierarchy. Its type on a meta-conceptual level is City with supertype Toponym. On the linguistic level, the conceptual proper name ID=3200 is realized in French as a prolexeme *Paris*, which has no aliases, but has a derivative *Parisien* - an inhabitant of *Paris* - that has a synonym *Parigot* - an inhabitant of *Paris* in French slang. These two lexemes are related by synonymy. Though synonymous, there are sociolinguistic differences between them as well as differences in the communication situations in which they are used. It should be noted that the prolexeme of the second lexeme *Parigot* is empty (ø), and the ID value of its conceptual proper name differs from the ID value of *Parisien.* On the level of instances, only one instance *Paris* marked as a masculine gender noun (M) in singular (S) corresponds to the prolexeme *Paris*, while four instances correspond to derivative *Parisien* defined by its inflective paradigm. The prolexeme *Paris* has as a derivative the relational adjective *parisien* with its own instances - they are not represented in Figure 2.

In Serbian, the prolexeme corresponding to the same conceptual name with ID=3200 is *Pariz*, and its alias is its Cyrillic recording *Париз*. Derivational processes in Serbian are more complex than in French. Besides the relational adjective *pariski*, and the name for a masculine inhabitant, *Parižanin*, the separate form exists for a feminine inhabitant, *Parižanka*. From inhabitant names, a relational adjective is derived *parižanski* (which is related to the inhabitants of *Paris*), and possessive adjectives *Parižaninov* (belonging to a *Parižanin*) and *Parižankin* (belonging to a *Parižanka*) - these derived forms are not represented in Figure 2. On the level of instances, the set of inflected forms corresponds to the prolexeme, its aliases, and all derived forms, the correspondence being established by appropriate regular expression.

It should be noted that in Serbian, the derivational level has itself two levels: on the first level are forms derived directly from the prolexeme or its aliases, while on the second level are forms that are being systematically produced from derived forms by the mechanism of structural derivation. As in Serbian an alternative name exists for the inhabitant of Paris, *Parizlija*, which corresponds to the French *Parigot*, its conceptual ID would also be 3201. Neither for German nor English do such forms exist.

5. Conclusion

The basic structure of a model of a multilingual relational dictionary of proper names based on a four-level ontology has been presented. It is envisaged that Unicode should be used for encoding, a relational database for the model implementation and XLM schema for data exchange. Such a database will be used in translation,

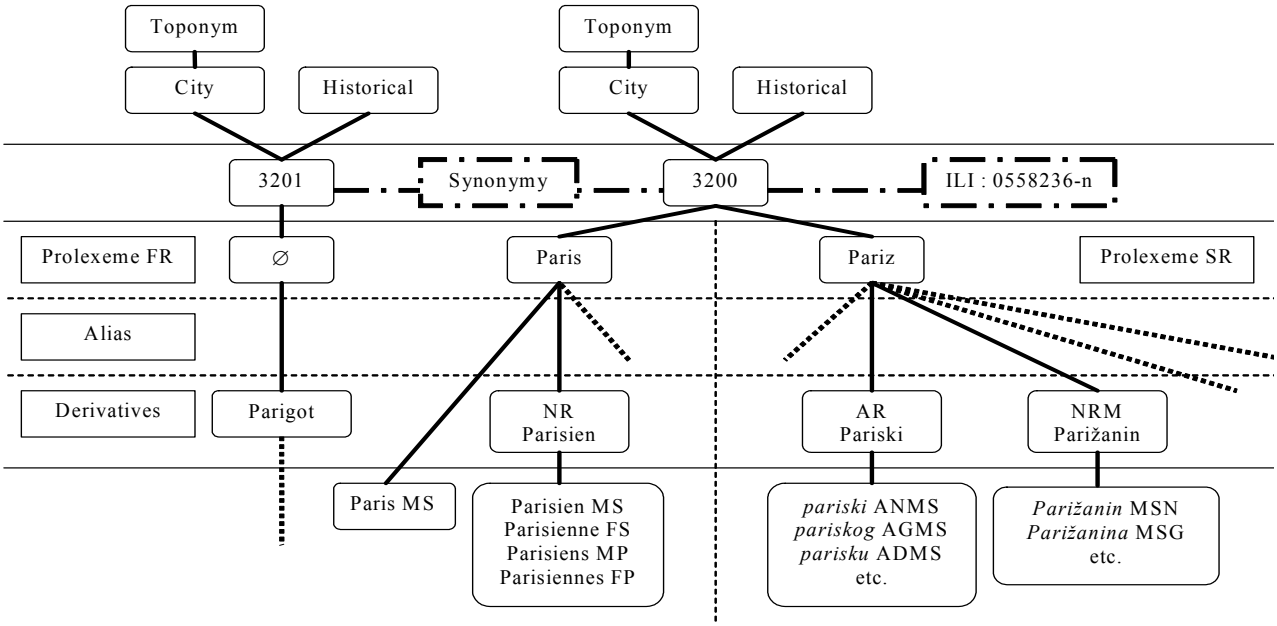

Figure 2. The concept of a proper name *Paris* in French and Serbian

information retrieval, text alignment, to mention just the main foreseen applications. When completed, the database should cover most of the European languages.

6. References

Coates-Stephens S. 1993. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. Hingham, MA: Kluwer Academic Publishers

Courtois, B.; and Silberztein M. 1990. Dictionnaires électroniques du français. *Langues française* 87: 11-22

Cucchiarini C., Daelemans W., Strik H. (2000), *Strengthening the Dutch Human Language Technology Infrastructure*, http://www.elda.fr/article48.html.

Grass T.; Maurel D.; and Piton O. 2002. Description of a multilingual database of proper names. In *Lecture Notes in Computer Science*, 2389: 137-140.

Gross, M. 1997. The Construction of Local Grammars. In Roche, E. and Schabes, Y. eds.: *Finite-State Language Processing*. 329-354. Cambridge, Mass: The MIT Press.

Gruber T. R. 1995. Toward Principles for the Design of Onthologies Used for Knowledge Sharing. *International Journal of Human-Computer Strudies* 43: 907-928

MacDonald D. 1996. Internal and external evidence in the identification and semantic categorisation of Proper Names. *Corpus Processing for Lexical Acquisition*, 21-39, Massachussetts Institute of Technology

Maurel, D. 2004. Les mots inconnus sont-ils des noms propres?. *JADT 2004*, Louvain-la-Neuve, Belgium. 776-784

Mel'cuk I. 1984-I, 1988-II, 1992-III. *Dictionnaire explicatif et combinatoire du français contemporain.* Les presses de l'Université de Montréal.

Mikheev A.; Moens M.; and Grover C. 1999. Named entity Recognition without Gazetteers, *EACL'99* Bergen, Norway: ACL June 1999. pp. 1-8

Sager J. C. 1990. *A Practical Course in Terminology Processing*, Amsterdam: John Benjamins

Tran M.; Maurel D.; and Savary A. 2005. Implantation d'un tri lexical respectant la particularité des noms propres, *Linguisticae Investigationes*. Forthcoming.

Vossen P. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.